Secure Data Marketplaces

Daniel Petri Rocha, B.Sc., Dr. Holger Kinkelin*, Filip Rezabek, M. Sc.*
*Chair of Network Architectures and Services, Department of Informatics
Technical University of Munich, Germany
Email: daniel.petri@tum.de, kinkelin@net.in.tum.de, frezabek@net.in.tum.de

***Abstract*—Big data powers the growing data economy. But critical data sets needed for research and development remain isolated. Those selling such data have no means of preventing copies from being created. This paper summarizes the building blocks required to construct a secure data market, where privacy and control are inherently built into the system despite large-scale information access remaining possible. Mechanisms for granting access to data as desired by the owner are described, enabling data to be leased without exposing it. Secured data processing techniques and blockchain technologies are suitable for assembling privacy-preserving data marketplaces.**

***Index Terms*—data market, data silo, blockchain**

1. Introduction

Data is valuable — the European Union's data economy alone is forecast at 550 billion Euros by 2025 [1]. The analysis of vast data collections fuels technologies that aid in the accurate and deep understanding of areas of societal importance, yielding, for instance, advancements in scientific research. Big data sets help uncover treatments for severe illnesses by studying the human genome [1]; healthcare centers can improve patient care with shared information [2]. However, trading such data often does not happen in practice, given that a separate entity with a copy could choose to redistribute it indiscriminately.

Therefore, enterprise information holders keep a monopoly on highly demanded data even though they know its value [3]. These data silos are a financial loss and liability source since a breach would leak business secrets or personally identifiable information without usage restrictions, artificially imposing a cap on the data's potential [3]. Instead, incentives should make organizations provide data to others in a discoverable and integrable manner [4].

A way to accomplish this without exposing sensitive information is by selling data as a good or service on secure data marketplaces. They offer the tools needed to create an additional revenue stream for data holders while ensuring the owner's data privacy and control [2]. Consumers use the market to locate and access the data they need without trusting a central authority. They can then work with it perpetually, only for a period or a limited number of times [3]. Interactions against the bought data occur with programs, e.g., running queries or machine learning algorithms to get a trained model back. Unlike downloading a file or tapping a stream, data is employed without risking it being cloned as it does not leave the holder's premises.

This paper is structured as follows: Section 2 describes where and how such technology is being applied or envisioned to be. Section 3 explains the desirable properties of data markets in conjunction with their motivation. Section 4 addresses what a data market consists of. Section 5 provides approaches to designing, implementing, and maintaining data marketplaces with the identified parts and properties. Section 6 concludes this proceeding.

2. Application Scenarios

Data marketplaces available today meet specific information demands. The subsections below analyze prominent data market applications and where to find them.

2.1. Ocean Market

The Ocean Protocol is a project attempting to supply an interface to simplify setting up data markets. Their open-source protocol provides the necessary infrastructure to give and withdraw paid data access. A transport company could increase revenue by using Ocean to deploy a data market for annotated dashcam footage they currently silo, which may be of interest for the computer vision models of the automakers engineering self-driving vehicles.

An exemplary Ocean-powered market is the Ocean Market[1], on which users pay with their crypto wallet. In return, they redeem a token for that asset, which can be considered a license of the original data set. Depending on the fine-grained permissions set by the data owner, this license is the ticket to data services such as downloading a copy or using it as input for an algorithm without revealing the underlying data. The provider can approve or deny programs to avoid privacy infringements. Ocean itself does not store any data: ownership corresponds to minting a non-fungible token on the Ethereum blockchain pointing to an external resource [3].

However, a decentralized identifier for the resource is stored on-chain, together with a separate document offering a metadata description to make it easily discoverable. For precise data control access, required credentials may be embedded in the metadata store, representing an additional type of identification besides token ownership. Using a role-based access control server whose implementation Ocean provides, capabilities around service consumption can be restricted.

1. https://market.oceanprotocol.com/

doi: 10.2313/NET-2022-11-1_09

Besides self-hosting a secure market, opening up siloed databases, or refining public data sets by making them effortlessly integrable and discoverable, an additional income stream can come from staking on data. Since Ocean Protocol data sets are tokenized and have value, prices can automatically be determined by an automated market maker — effectively transforming the data set into a cryptocurrency. The market maker dictates how expensive a token should be given its availability in a liquidity pool. The pool's liquidity is defined by the number of data tokens and cryptocurrencies such as Ethereum or Ocean (the organization's coin) it holds. An asset's cost increases as it is purchased and used, paying dividends to those providing liquidity. It decreases when data tokens are sold since the automated market maker derives the price from the assumption that the ratio of data tokens to, e.g., Ocean should stay constant at 50:50. Thus, a monetary incentive exists to curate data, which provisions liquidity [3]. An engineer, satisfied with the better rate of red traffic lights detected in their model after using a data set, may choose to stake it for profit. That, in turn, signals the market that the data set's quality is high.

2.2. Kara

Kara[2] is a secure market for medical data whose goal is to improve research and outcomes in medicine by recognizing that valuable health-related databases are siloed and offering a user-centric solution that lets patients share data themselves. The Oasis blockchain [5] fuels it.

After a doctor's visit, e.g., to perform a back of the eye scan, they let patients upload that image to a medical data cloud in exchange for cryptocurrency alongside a policy of use. An example guideline the patient provides could be to revoke their scan's use for commercial purposes or only grant research access for a certain number of years. They remain the data owner at all times instead of an intermediary company they may not trust.

Collaboration is fostered among doctors and scientists as the data units are inherently sharable due to a privacy-preserving architecture, ensuring the unencrypted scan can never be seen. Nonetheless, surveying the information remains possible, with applications including performing statistical analysis for genomics research and training machine learning models against the data set.

3. Properties of Data Markets

Transactions in a data marketplace occur between a data supplier, which owns a unit of information they are willing to sell, rent, or barter, and a consumer ready to enter the trade to obtain access [6]. As rational participants, a guarantee that a subset of the characteristics described below is enforced may interest both parties and the marketplace network to create a growing self-sustaining environment.

Data as (crypto-)currency A *currency* is a medium that facilitates trade. As with money, data is an asset belonging to an individual and can therefore be classified as a currency [2]. Personal data is part of how access to some online services remains free, being given up as payment. Data markets could enable users of these platforms to be remunerated for the use of their information in novel ways, an example being micro-deposits of cryptocurrency. In the Kara market, patients whose medical data played a role in training artificial intelligence models get to choose charities to which donations will be made. On Ocean, data is published as a non-fungible token from which tokens for access are created, similar to a cryptocurrency's initial coin offering. A user's crypto wallet then becomes, in effect, a data wallet for a stable commodity currency backed by sets of data [3]. Consequently, Ocean as a coin does not inherit other cryptocurrencies' fiat-like properties.

Discoverability Markets are only attractive to consumers if they can fulfill their data needs, which cannot happen if data sets are difficult to find. Incentives need to exist for sellers to describe their assets appropriately, with the market platform potentially being capable of blending multiple data sets based on metadata [4]. Suppose a customer wants to train their machine learning model on road signs, for instance. In that case, pictures from two separate collections, `EuropeanTrafficImages` and `AmericanTrafficImages`, could automatically be fed into the program, remunerating each seller.

Fairness An exchange should only occur if the seller and sellee concur with the transaction's commodity, policy, and price. A policy may define the data's extent to which it can be used: by whom, for what purposes, and for how long. Trade is fair if the buyer receives what they paid for, the policies are followed, and the seller is remunerated. Since either party involved can withdraw from the sale, fairness leads to both leaving empty-handed in that case [6].

Integrability Data may come from differing sources, contain missing or erroneous information, and be available in a format not directly usable by a consumer. Integrable data has been extracted, transformed, and cleansed. This time-consuming process makes it worth more than raw data. A market should provide incentives for sellers to prepare their assets in this handleable way [4].

Ownership Ownership must be kept track of publicly in a data market to preserve intellectual property rights. Howbeit, traditionally proprietorship of information is in the hands of silos in place of the individuals responsible for generating it. Data exhaust emanating from passive activities such as a purchase on a web store, interactions with smart sensors, and browsing history are monetizable yet do not enjoy the legal protection brought on by copyright laws for active data creation, e.g., writing an email [2]. They are covered by privacy laws instead. Data markets can help establish an economic model ascertaining people control their data through policies.

Provenance Knowing where data came from permits buyers and sellers to audit transactions better. The source may be a factor in determining a data set's quality: inferred data, for instance, is likelier incorrect [2].

Quality Information is prone to change. As it stales over time, its value decreases [2]. Therefore, data correctness is a factor in data markets if the price is automatically discovered.

Security Suppliers of data need to be assured that a leak can not occur. A fair trade in which no other party besides the buyer and seller can see the information (including

2. https://kara.cloud/

doi: 10.2313/NET-2022-11-1_09

intermediaries) is privacy-preserving [6]. However, piracy cannot be prevented if a dishonest buyer receives a full copy of the data. As a result, data escapes need to be prevented by only allowing compute access to sensitive information [3]. Still, the computation cannot occur on a remote machine set up by a cloud operator unless secured computing techniques are employed. Privacy issues and concerns riddle cloud providers. With clients not storing data locally, the attack surface is increased: they may be subject to having their virtual machines cloned or tampered with; audits are burdensome. The cloud provider may subcontract to third parties, making compliance with regulations hard since it is unclear whose responsibility and jurisdiction the data falls under.

Transparency Pricing transparency is necessary for exchanges involving a trade facilitator, i.e., a mediator that may host the marketplace platform. Buyers should be aware of the initial price and the terms of use of the transaction between the seller and the broker [6].

4. Building Blocks

In a secure data market, a seller must convince buyers that they have data the customer needs without revealing it at any point in the trade. Consumers, meanwhile, require assurances that their investment will return the desired results even though they do not know how valuable the data set is in advance [4]. That is fundamentally different from traditional product sales, as the partakers in the exchange are not dealing with a physical good. Instead, a service is provided to a data consumer in which they never get a copy of the information used to produce the final output [5]. Sophisticated technologies come into play to realize this paradigm shift.

4.1. Blockchain

Buying data off silos is problematic since they are in control of an entity that needs to be trusted not to modify it unfairly. A blockchain is a suitable data structure to permanently and irreversibly store the state of a database. Blockchains are immutable, i.e., entries are non-erasable and non-modifiable, meaning that once a transaction is added to the ledger, malicious actors cannot change its contents [7]. Additionally, they are decentrally run and managed, removing the need for an intermediator. As a component of secure markets, blockchains make the entire transaction history traceable, logging accesses and what buyers used it for. Health data usage, for instance, is required to comply with regulations. The chain's transparency lets them track ownership of, e.g., patient records and monitor whether the rules are followed since the ledger's nature ensures entries can only be added but not removed [6]. People that sold Kara their X-rays could see how and where their data is employed in the data economy.

4.2. Smart contracts

Smart contracts are programmable agreements executed on a blockchain-based architecture [8]. In data markets, contracting parties can algorithmically describe the terms of use of private data in the form of a policy [9], such as with whom providers can share banking data. An automatic market maker to establish the price of assets can also be implemented as a smart contract. The contract's code is cryptographically secured and automatically runs once agreed-to conditions are met, e.g., only starting training a supplied machine learning model after the funds have been received. A smart contract can thus act as a trusted trade intermediator in this context [5].

4.3. (Non-fungible) data tokens

Controlling access to data and maintaining intellectual property rights are tasks to be solved in secure data markets such as Kara and Ocean. A naive approach for managing access would be to issue a ticket for users that paid for a service. However, sharing the pass with numerous others, even those who should be barred from possessing one, would be trivial [3]. A mechanism to impose digital scarcity preventing the repeated spending of the same ticket is therefore needed. Blockchain architectures enable this through tokens, whose ownership is tracked and which can either be fungible or not.

A fungible token is identical to others of the same denomination. Holding a data token called `$LabeledTrafficImages`, for example, is a permit that gives the same functionality when using the data services as any other token of that instance. As an analogy, a 1€ coin has equal value to another.

On the other hand, a non-fungible token (NFT) is a digital deed for an asset —– like a collection of labeled traffic images —– that can be stored on the chain, conveying ownership over that property. The non-fungibility comes from the realization that data sets differ, as do physical belongings.

While NFTs could act as data tokens to solve the double-spending issue, the pictures likely interest more than one person. Therefore, the proprietor instead mints an NFT to represent possession of that asset's intellectual property and issues a limited number of licenses (`$LabeledTrafficImages` data tokens) at their discretion to the annotated photos.

Since data tokens can be transferred, a form of identification could additionally be required to redeem the service to combat unrestricted access by people without proper credentials [3].

4.4. Secured data processing

The data market component in which the final output is produced must be secured to prevent intellectual property rights violations and sensitive disclosures of personal digital information. A trusted execution environment offers this degree of protection by computing the result of the buyer's program in a figurative black box [5]. Decrypted data can never be interacted with from the outside by manipulating the data in enclaves, i.e., containers holding the confidential information to be processed and the instructions on how to perform the computation [10]. The application's address space is encrypted in memory [11] and decrypted by the computer's central processing unit.

In some data market architectures, the seller includes the decryption key in the smart contract alongside the use

doi: 10.2313/NET-2022-11-1_09

policy. The key is revealed once the provider's contract verifies that the customer's request satisfies the terms of use [9] [12]. The buyer's smart contract then performs the operations on the raw data inside the trusted execution environment. Alternatively, research and industry standardization efforts are underway [13] for techniques directly performing the computation on encrypted data, namely fully homomorphic encryption.

4.5. Program rewriting and verification

Ensuring policy compliance is an additional challenge requiring a separate module. The motivation of this component is to have a verifier mechanically determine that, given the buyer's program and the data owner's terms of use, the program will not disclose private information once executed. The output is a sound boolean value, meaning that if it certifies the policy is followed, proof of that fact is provided. In case the verifier can not assure that the conditions governing the data's use are met, the module could rewrite the program into one that does so [5].

Say a company wants to price a new product and purchases access to a data set about customers in their target market. They then write a program to query the mean income of students in Bavaria. Intuitively, this is different from asking the average salary of pupils enrolled at the Technical University of Munich with Alice as a first name. While the former question is broad enough to pass the verification step, the latter inquires about an individual and does not bode well with Alice's policy. However, if she is the only person in the data set, both queries are identical.

If removing Alice's record from the data set significantly affects the program's output, the query is said not to be differentially private enough. Differential privacy is a property that algorithms like deep learning models can fulfill that limits how much information concerning their inputs can be revealed [14]. Open-source implementations of differential privacy tools are emerging for general public use, contributing to the adoption of the technique in academia and the industry [15]. To accelerate its prevalence, systems exist that seamlessly integrate with current SQL databases and automatically rewrite queries enforcing differential privacy [16].

5. Architecture

With a secure data market's components and properties now identified, a sample architecture for its realization is provided next. A distinction is made between the stage in which data is added to the market and the one where it is acquired, as they run asynchronously.

1) **Publish step**
a) A seller publishes an encrypted data set to a cloud storage service accessible via a URI that compute jobs may need [9].
b) They provision a smart contract that mints an ERC721[3]-compliant non-fungible token on the Ethereum blockchain pointing to their service [3], claiming themselves as the intellectual property rights holder.
c) The decryption key is secretly kept in the contract with constraints such as the use policy [9].
d) In the desired quantity, the seller's smart contract also mints a pool of ERC20[3] data tokens (licenses) for utilizing the service [3].
2) **Consume step**
a) After identifying relevant data for their purposes in a market's front-end interface, a shopper pays for a data token and writes a smart contract with the code they want to run using the seller's data as input.
b) Once executed, the contract transfers a data token to the seller's crypto wallet as a request to rent access to the data [3].
c) The seller's contract verifies that performing the request will be privacy-preserving (rewriting the request or withdrawing from the sale if necessary) and returns the decryption key for use inside a trusted execution environment [9].
d) The buyer's smart contract runs securely in a trusted execution environment, returning only the computation result.

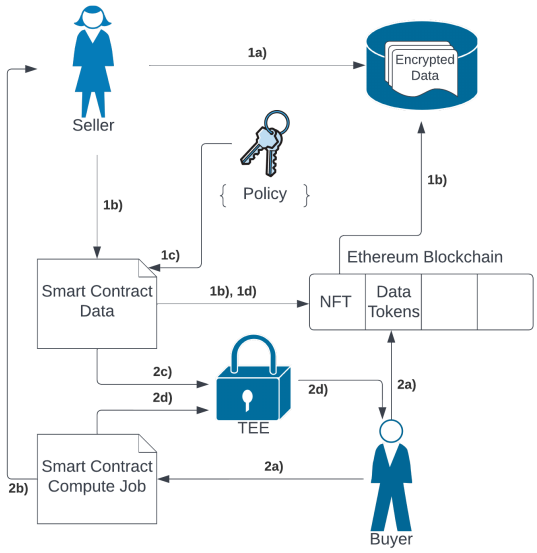

Figure 1: Data sale transaction as outlined in Section 5.

6. Conclusion

In this paper, we established that even though value can be extracted from raw data, little incentive exists for people to curate it in a way that makes it accessible and usable by others. Existing secure data markets, e.g., Kara and Ocean, encourage such behavior while assuring user privacy and control over their information is paramount. Blockchain technologies are fitting in tackling such a task as they enable data trading in a controlled fashion. With smart contracts facilitating data transactions, ensuring adherence to terms of use, and automatic price determination, data assets turn into cryptocurrencies supported by real-world applications.

3. "Ethereum Request for Comment": technical proposal for a standard

doi: 10.2313/NET-2022-11-1_09

References

[1] European Commission, Directorate-General for Communications Networks, Content and Technology, Cattaneo, G., Micheletti, G., Glennon, M., et al., *The European Data Market Monitoring Tool: Key Facts & Figures, First Policy Conclusions, Data Landscape and Quantified Stories: d2.9 Final Study Report*. Publications Office, 2020.

[2] C. Gates and P. Matthews, "Data Is the New Currency," in *Proceedings of the 2014 New Security Paradigms Workshop*, ser. NSPW '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 105–116. [Online]. Available: https://doi.org/10.1145/2683467.2683477

[3] "Tools for the Web3 Data Economy," https://oceanprotocol.com/tech-whitepaper.pdf, Ocean Protocol Foundation with BigchainDB GmbH, Tech. Rep., 2022, last accessed on 2022/05/22.

[4] R. C. Fernandez, P. Subramaniam, and M. J. Franklin, "Data Market Platforms: Trading Data Assets to Solve Data Problems," *Proc. VLDB Endow.*, vol. 13, no. 12, p. 1933–1947, Jul. 2020. [Online]. Available: https://doi.org/10.14778/3407790.3407800

[5] N. Johnson, "Building a Secure Data Market on Blockchain." Burlingame, CA: USENIX Association, Jan. 2019.

[6] P. Banerjee and S. Ruj, "Blockchain Enabled Data Marketplace – Design and Challenges," https://arxiv.org/abs/1811.11462, 2018.

[7] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," Dec. 2008, last accessed on 2022/06/04. [Online]. Available: https://bitcoin.org/bitcoin.pdf

[8] V. Buterin, "A Next Generation Smart Contract & Decentralized Application Platform," 2015.

[9] N. Hynes, D. Dao, D. Yan, R. Cheng, and D. Song, "A Demonstration of Sterling: A Privacy-Preserving Data Marketplace," *Proc. VLDB Endow.*, vol. 11, no. 12, p. 2086–2089, Aug. 2018. [Online]. Available: https://doi.org/10.14778/3229863.3236266

[10] V. Costan and S. Devadas, "Intel SGX explained," *IACR Cryptol. ePrint Arch.*, p. 86, 2016. [Online]. Available: http://eprint.iacr.org/2016/086

[11] D. Lee, D. Kohlbrenner, S. Shinde, K. Asanović, and D. Song, "Keystone: An Open Framework for Architecting Trusted Execution Environments," in *Proceedings of the Fifteenth European Conference on Computer Systems*, ser. EuroSys '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3342195.3387532

[12] D. Dao, D. Alistarh, C. Musat, and C. Zhang, "DataBright: Towards a Global Exchange for Decentralized Data Ownership and Trusted Computation," 2018. [Online]. Available: https://arxiv.org/abs/1802.04780

[13] M. Albrecht, M. Chase, H. Chen, J. Ding, S. Goldwasser, S. Gorbunov, S. Halevi, J. Hoffstein, K. Laine, K. Lauter, S. Lokam, D. Micciancio, D. Moody, T. Morrison, A. Sahai, and V. Vaikuntanathan, "Homomorphic Encryption Security Standard," Toronto, Canada, Tech. Rep., November 2018.

[14] N. Carlini, C. Liu, U. Erlingsson, J. Kos, and D. Song, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," in *Proceedings of the 28th USENIX Conference on Security Symposium*, ser. SEC'19. USA: USENIX Association, 2019, p. 267–284.

[15] "The OpenDP White Paper," https://opendp.org/, OpenDP, Tech. Rep., May 2020, last accessed on 2022/06/09.

[16] N. M. Johnson, J. P. Near, J. M. Hellerstein, and D. Song, "Chorus: Differential Privacy via Query Rewriting," *CoRR*, vol. abs/1809.07750, 2018. [Online]. Available: http://arxiv.org/abs/1809.07750